PROPER SCORES FOR PROBABILITY FORECASTERS[1]

and

DETERMINING SUBJECTIVE PROBABILITIES

BY SEQUENTIAL CHOICES[1]

by

Arlo D. Hendrickson and Robert J. Buehler

Technical Report No. 158

June 1971

University of Minnesota

Minneapolis, Minnesota

[1] This research was supported by National Science Foundation Grant GP-9556.

PROPER SCORES FOR PROBABILITY FORECASTERS[1]

by

Arlo D. Hendrickson and Robert J. Buehler

Virginia Polytechnic Institute and University of Minnesota

ABSTRACT

A probability forecaster is asked to give a density p of a random variable ω. In return he gets a reward (or score) depending on p and on a subsequently observed value of ω. A scoring rule is called proper if the expected score is maximized when the true density is chosen. The present paper uses convex analysis to generalize McCarthy's characterization of proper scoring rules.

[1]Research supported by NSF Grant GP-9556.

1. <u>Introduction and Summary</u>. Let $(\Omega, \mathcal{G}, \mu)$ be a measure space and let $\mathcal{P}$ be a convex class of probability densities with respect to the measure μ. A scoring rule f is a mapping from $\mathcal{P}$ into the class $\mathcal{L}$ of random variables on Ω. Assume a forecaster has knowledge of a probability density $p \in \mathcal{P}$, and is to receive the score (or actual payment) $f(p)$ for his disclosure of p. Since $f(p)$ is a random variable, the score depends on the outcome of the experiment $\omega \in \Omega$. The score f has been called <u>proper</u> if

$$(1) \qquad E_p(f(p)) \geq E_p(f(q)) \text{ for all } p, q \in \mathcal{P}$$

where $E_p(\cdot)$ is the mathematical expectation with respect to the density p. If (1) holds, then the forecaster will maximize his expected score with respect to p by disclosing this density p. To avoid difficulties in (1) we will assume $E_p(f(q))$ exists and is finite.

The first suggested use of a scoring rule was apparently by Brier (1950) in connection with weather forecasting. The independent work of Good (1952) explicitly considered condition (1). For more recent work, see for example de Finetti (1962), Winkler (1969), Savage (1970), and Staël von Holstein (1970). The latter notes that scoring rules have also been called payoff or reward or incentive functions, and gives an excellent bibliography listing 133 items.

While the above context in which p is known to the forecaster is adequate for our purposes, other points of view are possible. For example the cumulative score of any forecaster may be used as a measure of his forecasting ability; or the stated p may be regarded as defining a subjective probability. But the present paper is concerned only with the purely mathematical problems connected with the characterization of functions f satisfying (1) (or strictly satisfying (1)). Our main result is Theorem 3.1, which modifies and generalizes a theorem of McCarthy (1956), using a generalization of Rockafellar's (1970) definition of subgradient.

Theorems 4.2, 4.3, and 4.4 give additional conditions to ensure that there exists an f satisfying the requirements of Theorem 3.1. The necessary preliminary definitions and theorems are given in Section 2.

2. Some Concepts of Convex Analysis. The space $\mathcal{L}$ of random variables on $(\Omega, \mathcal{G})$ is a vector space with an inner product defined whenever it exists by

$$(2) \qquad p \cdot q = \int p(\omega)q(\omega)d\mu(\omega).$$

Let $\mathcal{L}_1 = \mathcal{L}_1(\mathcal{P})$ be the set of all $q \in \mathcal{L}$ such that $p \cdot q$ is defined for all $p \in \mathcal{P}$. The range of a scoring rule f defined on $\mathcal{P}$ is assumed to be contained in $\mathcal{L}_1(\mathcal{P})$. The following relation is crucial for applying convex analysis to studying (1):

$$(3) \qquad E_p(q) = p \cdot q \qquad \text{if} \quad p \in \mathcal{P} .$$

For given f, the expected score H is defined on $\mathcal{P}$ by

$$(4) \qquad H(p) = p \cdot f(p)$$

and condition (1) is equivalent to

$$(5) \qquad H(p) \geq p \cdot f(q) \qquad \text{for all} \quad p, q \in \mathcal{P} .$$

The condition that f be strictly proper is

$$(6) \qquad H(p) > p \cdot f(q) \qquad \text{if} \quad p \neq q.$$

The following is a generalization of Rockafellar's (1970) definition of subgradient to the infinite-dimensional case:

Definition 2.1. If H is defined on a convex set $D \subset \mathcal{L}$ and if there exists $q \in D$ and $q^* \in \mathcal{L}_1(D)$ such that

$$(7) \qquad H(p) \geq (p-q)\cdot q^* + H(q) \qquad \text{for all } p \in D$$

then q^* is a subgradient of H at q (relative to D).

It can be shown in the Euclidean case that a subgradient of a convex function H is unique and equal to the gradient at every point where H is a differentiable (Rockafellar, (1970) Theorem 25.1).

Theorem 2.1. If H has a subgradient q^* at each point q in a convex set D, then H is convex on D.

Proof: For any $p, q \in D$, let p_1^* be a subgradient of H at $p_1 = (1-\lambda)p + \lambda q$. Then $H(p) \geq (p-p_1)\cdot p_1^* + H(p_1)$ and $H(q) \geq (q-p_1)\cdot p_1^* + H(p_1)$. It follows that $(1-\lambda)H(p) + \lambda H(q) \geq H(p_1)$.

The following is a variant of Euler's theorem.

Theorem 2.2. If H is homogeneous of degree r on a convex cone D and has a subgradient $q^* \in \mathcal{L}_1(D)$ for some $q \in D$, then $rH(q) = q\cdot q^*$.

Proof: The inequality $\lambda^r H(q) = H(\lambda q) \geq (\lambda q - q)\cdot q^* + H(q)$ implies

$$\frac{\lambda^r - 1}{\lambda - 1} H(q) \geq q\cdot q^* \qquad \text{if } \lambda > 1,$$

with the reverse inequality if $0 < \lambda < 1$. Thus $rH(q) = q\cdot q^*$.

It can be shown further that if q^* is a subgradient at q of H in Theorem 2.2, then $(\lambda q)^* = \lambda^{r-1} q^*$ is a subgradient of H at λq for all $\lambda > 0$.

In the sequel we will use the term "homogeneous" to mean "homogeneous of the first degree." If H is homogeneous on a convex set C, then by letting $H(\lambda p) = \lambda H(p)$ we can always extend the domain of H to the convex cone $D = \{\lambda p: p \in C, \lambda > 0\}$.

3. McCarthy's Theorem. McCarthy (1956) stated without proof a characterization of proper scoring rules for the case when $\mathcal{P}$ is the class of discrete distributions on a finite set Ω. Our Theorem 3.1 applies to more general $\mathcal{P}$ and distinguishes between strict and non-strict inequalities.

Theorem 3.1. A scoring rule f mapping $\mathcal{P}$ into $\mathcal{L}_1$ satisfies (1) [strictly] iff there exists a function H defined on $D = \{\lambda p: p \in \mathcal{P}, \lambda > 0\}$ which is (a) homogeneous, (b) convex [strictly convex on $\mathcal{P}$], and (c) such that $f(p)$ is a subgradient of H relative to D at p for all $p \in \mathcal{P}$. The function H satisfies $H(\lambda p) = \lambda p \cdot f(p)$.

Proof: Assuming (1) holds, define $H(\lambda p) = \lambda p \cdot f(p)$. Using (1)

$$H(\lambda p) \geq \lambda p \cdot f(q) = (\lambda p - q) \cdot f(q) + H(q)$$

for all $p, q \in \mathcal{P}$, $\lambda > 0$, which establishes (c). Finally (b) follows from Theorem 2.1.

Conversely, (a, b, c) imply by Theorem 2.2 that $H(\lambda p) = \lambda p \cdot f(p)$, and substituting this into the subgradient inequality gives the desired result (1).

Strict inequality in (1) is equivalent to no subgradient of H at p being a subgradient of H at q, if $p, q \in \mathcal{P}$, $p \neq q$. This is equivalent to H being strictly convex on $\mathcal{P}$.

Example 3.1. A familiar example is the logarithmic score suggested by Good (1952) for the binomial case. In the general case we put

(8) $\qquad f(p) = \ln p.$

A well known inequality shows that f is strictly proper. If μ is finite, then $H(\lambda p) = \lambda p \cdot f(p)$ is finite for $\lambda p \in \mathcal{L}_2^+$ where $\mathcal{L}_2^+ = \{q: q(\omega) \geq 0$ for all $\omega \in \Omega, \int q^2 d\mu < \infty\}$, and H is continuous with respect to the $\mathcal{L}_2$ norm $\|\cdot\|$ defined by the inner product (2).

In the finite discrete case let the density $p(\omega)$ be replaced by a vector p of probabilities p_j. For this $p \in \mathcal{P}$, $H(p) = \Sigma p_j \ln p_j$, but for $q = \lambda p \in D$, $H(q) = \lambda \Sigma p_j \ln p_j = \Sigma q_j \ln(q_j/\Sigma q_k)$. Marschak (1960), p. 97, attempted to show that the logarithmic score gave a counterexample to McCarthy's theorem, erroneously considering the gradient of $\Sigma p_j \ln p_j$ rather than of $\Sigma q_j \ln(q_j/\Sigma q_k)$. A proper understanding of the theorem requires a clear distinction between $\mathcal{P}$ and D not explicitly stated in McCarthy's paper.

If we wish to define f on D as well as on $\mathcal{P}$, then a natural choice is $f(\lambda p) = f(p)$. In particular for the logarithmic case $f(q) = \ln(q/\int q d\mu)$. Unlike $\ln q$, this $f(q)$ is a subgradient of H for all $q \in D$.

Example 3.2. For $\mathcal{P} \subset \mathcal{L}_2(\mu)$, the "quadratic" score $f(p) = 2p - \|p\|^2$ (Brier (1950), de Finetti (1962)) is strictly proper. $H(\lambda p) = \lambda\|p\|^2$.

Example 3.3. For $\mathcal{P} \subset \mathcal{L}_2(\mu)$, the "spherical" score $f(p) = p/\|p\|$ is strictly proper. $H(\lambda p) = \lambda\|p\|$.

4. Expected Score Functions. It might be asked what class of homogeneous and convex functions on D satisfy the additional requirement of Theorem 3.1 of having subgradients relative to D at each point in $\mathcal{P}$. The following is an example of a function which has no subgradients and yet is homogeneous and convex.

Example 4.1. Let $\mathcal{P}$ be the class of continuous, bounded densities ($\sup_{\omega} p(\omega) < \infty$) on (R, β, μ) where μ is Lebesgue measure and β consists of the Borel sets. Define $H(p) = \sup_{\omega} p(\omega)$. Then H is clearly convex on $\mathcal{P}$. However, H is neither continuous at any $p \in \mathcal{P}$ (with respect to $\|p\|$) nor does H have a subgradient for any $p \in \mathcal{P}$.

Let $\mathcal{H} \subset \mathcal{L}$ be a Hilbert space, R the real numbers, and let $\mathcal{H} \times R$ have the usual product topology and inner product. $\mathcal{H}$ can be taken to be the smallest closed subspace of $\mathcal{L}_2(\mu)$ containing $\mathcal{P}$, where $\mathcal{P} \subset \mathcal{L}_2(\mu)$. If $D \subset \mathcal{H}$ is the convex domain of a real-valued function H, then the epigraph of H, epi $(H) \subset \mathcal{H} \times R$, is the set $\{(p, \alpha) : \alpha \geq H(p),\ p \in D\}$. H is a convex function iff epi (H) is a convex set.

The following is a partial converse to Theorem 2.1.

Theorem 4.1. Let D be a convex set in $\mathcal{H}$ whose interior is nonempty. Let H be a convex function on D which is continuous at a point $p \in \text{int}(D)$. Then H has a subgradient $q^* \in \mathcal{H}$ at each point $q \in \text{int}(D)$.

Proof: The assumptions imply epi (H) is a convex subset of the Hilbert space $\mathcal{H} \times R$ whose interior is nonempty. If this is satisfied then epi (H) has a closed hyperplane of support through each of its boundary points. (See for example Valentine (1964) Theorems 2.15 and 4.1.) The supporting hyperplane at the boundary point $(q, H(q))$ is seen to give one of the following inequalities for some $q^* \in \mathcal{H}$:

$$(9) \qquad H(p) \geq (p-q) \cdot q^* + H(q) \quad \text{for all } p \in D,$$

or

$$(10) \qquad q \cdot q^* \geq p \cdot q^* \quad \text{for all } p \in D.$$

Clearly, (10) is satisfied only if $q \in \text{bdry}(D)$. Hence (9) is satisfied if $q \in \text{int}(D)$.

Theorem 4.2. If the set of densities $D \subset \mathcal{H}$ is a convex set and if H is convex and homogeneous on D and continuous at a point p in the interior of D, then there exists f such that conditions (4) and (5) hold on the interior of D. The range of f may be taken in $\mathcal{H}$.

Proof: Whenever $p \in \text{int}(D)$, apply Theorem 4.1 and let $f(p)$ be a subgradient of H at p. The proof follows from Theorem 2.2.

The following theorems give equivalent conditions on f for continuity conditions on H. We assume the range of f is in $\mathcal{H}$.

Theorem 4.3. If D is a convex cone in $\mathcal{H}$ whose interior is nonempty and if H and f satisfy (4) and (5) on D, then H is continuous at $p \in \text{int}(D)$ iff there exists a neighborhood of p on which $\|f(\cdot)\|$ is bounded.

Proof: Let $p, p_n \in D$, $\|p_n - p\| \to 0$ as $n \to \infty$. Let $q_n = f(p_n)/\|f(p_n)\|^2$. Then $H(p_n + q_n) \geq (p_n + q_n)\cdot f(p_n) = H(p_n) + 1$. Thus, if H is continuous at p, we cannot have $\|q_n\| \to 0$. Hence $\|f(\cdot)\|$ is bounded on a neighborhood of p.

Conversely, if $\|f(\cdot)\|$ is bounded on a neighborhood of p then by the Cauchy-Schwarz inequality $(p_n - p)\cdot f(p_n) \to 0$ if $\|p_n - p\| \to 0$. This implies $\overline{\lim}\, H(p_n) = \overline{\lim}\, p\cdot f(p_n) \leq H(p)$. Also $\underline{\lim}\, H(p_n) \geq \lim p_n \cdot f(p) = H(p)$. Hence, if $\|p_n - p\| \to 0$ then $H(p_n) \to H(p)$.

We will now assume that f is defined on the convex cone D such that

(11) $\quad f(\lambda p) = f(p) \quad$ if $p \in D$, $\lambda > 0$.

This condition, although natural, is not necessary because the homogeneous function H may have several subgradients at any $p \in D$.

Corollary 4.1. If f and H satisfy (4), (5), and (11) on a convex cone $D \subset \mathcal{H}$ then H is continuous on the interior of D iff $\|f(\cdot)\|$ is bounded on every closed set contained in D.

Proof: Since $f(\lambda p) = f(p)$ if $\lambda > 0$, $\|f(\cdot)\|$ is bounded on every closed set contained in D is equivalent to $\|f(\cdot)\|$ bounded on every compact set in D, which is equivalent to the requirement of Theorem 4.3 that $\|f(\cdot)\|$ be locally bounded at each point $p \in \text{int}(D)$.

Theorem 4.4. If H and f satisfy conditions (4), (5), and (11) for all points in a Hilbert space $\mathcal{H}$, then the following are equivalent:

(i) H is continuous;

(ii) H is bounded on the sphere $\{p \in \mathcal{H} : \|p\| = 1\}$;

(iii) $\|f\|$ is bounded.

Proof: We need only show (ii) implies (iii). This follows from $H(f(q)/\|f(q)\|) \geq (f(q)/\|f(q)\|) \cdot f(q) = \|f(q)\|$.

REFERENCES

[1] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review **78** 1-3.

[2] De Finetti, B. (1962). Does it make sense to speak of "good probability appraisers?" In Good (1962), 357-364.

[3] Good, I. J. (1952). Rational decisions. J. Royal Stat. Soc. Ser. B. **14** 357-364.

[4] __________ (ed.)(1962). The Scientist Speculates. Basic Books, New York.

[5] McCarthy, J. (1956). Measures of the value of information. Proc. Nat. Acad. Sci. **42** 654-655.

[6] Marschak, J. (1960). Remarks on the economics of information. Contributions to Scientific Research in Management. Univ. of California Press, Berkeley.

[7] Rockafellar, R. T. (1970). Convex Analysis. Princeton University Press, Princeton, New Jersey.

[8] Savage, L. J. (1970). The elicitation of personal probabilities and expectations. Yale University. Unpublished.

[9] Staël von Holstein, C.-A. S. (1970). Assessment and Evaluation of Subjective Probability Distributions. The Economic Research Institute at the Stockholm School of Economics, Stockholm.

[10] Valentine, F. A. (1964). Convex Sets. McGraw-Hill, New York.

[11] Winkler, R. L. (1969). Scoring rules and the evaluation of probability assessors. J. Amer. Statist. Assoc. **64** 1073-1077.

DETERMINING SUBJECTIVE PROBABILITIES

BY SEQUENTIAL CHOICES[1]

by

Arlo D. Hendrickson and Robert J. Buehler

Virginia Polytechnic Institute and University of Minnesota

ABSTRACT

Let A denote an uncertain outcome whose subjective probability is to be determined. A scheme is described wherein the subject is offered a sequence of choices of prospects. At step n, a value r $(0 < r < 1)$ is determined by previous responses, and Prospect A is a reward of $g(r)$ if A occurs, while Prospect B is a reward of $g(r)$ if B_r occurs, where B_r has known probability r. A characterization is given of those rewards which encourage honest responses. The results are related to the method of "score" or "payoff" functions for determining subjective probabilities.

[1]Research supported by National Science Foundation Grant GP-9556.

1. Introduction.

Definitions of subjective probability may be given in terms of preferences or choices. For example, in the translation by Kyburg and Smokler [11, p. 57] of Borel [2] we find:

> Paul claims that it will rain tomorrow; I agree that we are in accord on the precise meaning of this claim and I offer him the choice of receiving 100 francs if he is correct or 100 francs if he receives a 5 or a 6 in a throw of dice. In the second case the probability of receiving 100 francs is one third; if he then prefers to receive 100 francs if his meteorological prediction is correct, it is because he attributes to this prediction a probability superior to one third. The same method can be applied to all verifiable judgements; it allows a numerical evaluation of probabilities with a precision quite comparable to that with which one evaluates prices.

A similar example involving probabilities that eggs are good is given by Savage [14, p. 28].

In Borel's example the subject's choice fails to determine a unique value of the subjective probability of rain but tells us only whether it belongs to the interval $0 \leq p \leq \frac{1}{3}$ or to the interval $\frac{1}{3} \leq p \leq 1$. The present paper is concerned with procedures which determine subjective probabilities uniquely or else to a preassigned accuracy. Two ways of doing this are considered. The first is to present the subject with a choice of more than two alternatives. In Section 2 we show the relationship between this approach and the method of "score" or "payoff" functions. The second method is simply to repeat the binary choices of Borel's example sequentially with the prospects at each stage determined by previous choices. In Section 3 we show that the rewards in such sequential methods must be chosen with care for otherwise a dishonest answer may actually be advantageous to the subject. For a particular sequential scheme we characterize the rewards which "encourage honesty." In

Section 4 we consider how it is possible to modify the procedures so as to avoid any assumptions (such as linearity) of the subject's utility function.

2. Choices and Score Functions.

The main ideas of this section have evolved through the work of de Finetti, Savage, and others (as indicated). We give the essentials as background for Section 3. For a more detailed account, see Savage [15].

Let R be any uncertain but verifiable outcome, such as "rain tomorrow," and let p denote a subjective value of $P(R)$. Let Prospect A be a reward of one dollar if A occurs (and nothing otherwise) and let Prospect B be a reward of one dollar if H occurs (H = heads on the toss of a coin) and nothing otherwise. We take it to be axiomatic that $P(H) \leq p \leq 1$ if Prospect A is chosen and $0 \leq p \leq P(H)$ if Prospect B is chosen. If we agree that $P(H) = \frac{1}{2}$ and if the subject has a linear utility function (as we shall assume throughout Sections 2 and 3), then Prospect B is equivalent to a reward of $0.50 whether or not A occurs. This suggests a more general procedure in which a subject may choose one of n prospects $A_1, A_2, \ldots, A_n$, where A_j is a reward of a_j if R fails to occur or a reward of $a_j + b_j$ if R occurs. If we presume that the subject makes that choice which maximizes his utility, calculated using his subjective probability p, then the choice A_i implies that p satisfies

$$(2.1) \qquad a_i + b_i p \geq a_j + b_j p, \quad \text{for } j = 1,2,\ldots, n.$$

Thus when $\{a_j, b_j\}$ are suitably chosen, p is known to lie in one of

n intervals; and by increasing the number n of prospects we may determine p to any required accuracy.

The function $H_n(p) = \sup_j(a_j + b_j p)$ is convex and piecewise linear. For a continuous analog of the case of finite n, let $H_n(p)$ be replaced by any strictly convex $H(p)$ having a unique supporting straight line $L_{p'}(p)$, tangent to H at $(p', H(p'))$, for each $0 \leq p' \leq 1$. If $a_{p'} = L_{p'}(0)$, $a_{p'} + b_{p'} = L_{p'}(1)$, then $L_{p'}(p) = a_{p'} + b_{p'}p$. The subject may be given a continuous choice of prospects: for any chosen value $0 \leq p' \leq 1$ he receives $a_{p'}$ if R fails to occur or $a_{p'} + b_{p'}$ if R occurs. The choice p' has utility $L_{p'}(p)$, and from the strict convexity of H it is easily seen that

$$(2.2) \qquad H(p) = L_p(p) \geq L_{p'}(p) \quad \text{for all } p, p',$$

with equality only when $p' = p$. Since utility is maximized by the choice $p' = p$, it seems reasonable to define the individual's subjective probability to be the chosen value p'. This definition has in fact been proposed and studied by Savage [15, Sec. 8].

Occasionally it may be of interest to consider functions $H(p)$ which have linear segments (that is "flat spots," so that H is not strictly convex), or "corners," where there is more than one line of support. In the latter case we may agree that there are different choices all of which define the same value of the subjective probability, namely the abscissa of the corner. In the former case, a choice will correspond to a straight line supporting H over a range of values, and we will say that the subjective probability by definition belongs to that range. This in fact is the case which is considered in Section 3 below.

The case of a vector $(p_1, \ldots, p_k)$ of probabilities corresponding to k disjoint and exhaustive outcomes $R_1, \ldots, R_k$ can be treated similarly. In general we may wish to consider a "payoff" or "score" function $f_j(p')$ such that the subject receives $f_j(p')$ when p' is his stated probability vector and outcome R_j occurs. The utility to the subject of stating the vector p' is $\Sigma p_j f_j(p')$. The score function has been called proper if

$$\Sigma p_j f_j(p) \geq \Sigma p_j f_j(p') \quad \text{for all } p, p'. \tag{2.3}$$

If equality holds only when $p = p'$, then $f_j(p)$ is called strictly proper, and a stated p' can be defined to be the subjective probability vector.

Score functions were first proposed by Brier [5] and independently by Good [7], who introduced condition (2.3). For more recent work, see for example [1, 6, 13, 15, 16, 19, 21, 22]. Actual experimental results will be found in [18] and [20]. A characterization of proper scores given by McCarthy [12] has been generalized by Hendrickson and Buehler [10].

3. Determining Subjective Probabilities by Sequential Choices.

In this section we will consider a number of sequential procedures for determining subjective probability. The subject is offered a sequence of choices with the prospects at each stage depending on previous choices. First we will indicate a number of difficulties, and then we will describe a satisfactory procedure.

3.1 A procedure suggested by Borel.

In Borel [4; Chapter 3, Sec. 9], we find an extension of the idea mentioned in Section 1 above:

The method of auction sale often clarifies a buyer's exact evaluation of the worth of an object or a building offered for sale, for he stops bidding when the limit he has set himself has been reached. A similar method may be used, if Peter will accept it, to compel Peter to reveal precisely his evaluation of a certain probability. Let us return to the case where Peter is a consultant doctor who has been able to evaluate a patient's chances of recovery. We propose to find out whether he evaluates these chances at more than 50 per cent. We choose a contingent event with an exact 50 per cent probability, such as the game of heads or tails and we offer Peter an important gift or an intangible reward of considerable value to him and give him the choice between the following two eventualities: either he will receive the gift if the patient recovers, or he will receive the gift if a tossed coin turns up tails. He is clearly interested in choosing the eventuality with the higher probability in his opinion. He will choose the recovery of the patient if he considers the probability of this recovery to be above 50 per cent. If, on the contrary, he chooses the game of heads or tails, that will prove to us that he evaluates the probability of the recovery at less than 50 per cent. We may then repeat the test, using an eventuality with a 49 per cent probability. We may, for instance, with several decks of cards, whose backs are similar, make up a pack of 100 cards, 49 of which are red and 51 black. The probability of extracting a red card from this pack spread on the table after shuffling is 49 per cent or 0.49. If Peter prefers this probability to that of the case of recovery, it is because he evaluates the latter at less than 0.49. We may continue until Peter chooses the probability of the recovery when the other probability is only 0.43, whereas he had preferred the probability 0.44. We shall conclude that his true evaluation of the probability of the recovery is between 0.43 and 0.44. But, of course, true evaluation does not mean exact evaluation, since Peter is not infallible. Even if he is very skillful, it is quite doubtful that he can distinguish with certainty between probabilities as close together as 0.43 and 0.44. That is why it would be futile to seek a more exact decimal by diminishing the successive probabilities by a thousandth instead of a hundredth.

In actual practice it is relevant to know whether Peter is apprised of the rules of the questioning scheme. If he is not, then the possibility that he may guess the rules confuses the picture. Therefore we suppose the rules are spelled out in advance. To formalize Borel's suggested procedure, we will attach a utility to each prize, putting $g(q)$ equal to the utility value when the outcome of known probability has probability

q. Let A, B denote respectively the choice by Peter of the prospect involving the patient's recovery and that involving the outcome having known probability. Our understanding of the procedure is that the response sequence will have either the form BBB...BA or AAA...AB, where for example the response BBBBBBBA implies that the probability of recovery lies between 0.43 and 0.44. If we let p denote Peter's subjective value of the probability of recovery, then with the assumption of a linear utility function we find, for example:

Response	"Honest range"	Utility
AAB	$0.51 \leq p \leq 0.52$	$U_{AAB}(p) = pg(.50) + pg(.51) + 0.52g(.52)$
AB	$0.50 \leq p \leq 0.51$	$U_{AB}(p) = pg(.50) + 0.51g(.51)$
BA	$0.49 \leq p \leq 0.50$	$U_{BA}(p) = 0.50g(.50) + pg(.49)$
BBA	$0.48 \leq p \leq 0.49$	$U_{BBA}(p) = 0.50g(.50) + 0.49g(.49) + pg(.48)$

Here the "honest range" is the interval of p values which presumably contains the actual subjective value when the corresponding response is given. In accordance with the discussion in Section 2 above, this will be the case if and only if the response maximizes the utility over that range. But this is not the case. For example $U_{AAB}(p) = U_{AB}(p)$ should imply $p = 0.51$, but in fact implies $p = 0.51 - 0.52g(.52)/g(.51)$. Thus there is no (non-zero) choice of prizes $g(q)$ such that each response maximizes the utility over each appropriate "honest range." The essential reason for this is simply that in many cases it is more profitable for Peter to prolong the sequence, in order to obtain more prospects of prizes, than to give an "honest" answer.

3.2 An alternative procedure.

In view of the difficulties with the preceding method, we will now describe an alternative procedure. The present section will give the rules, and the following sections the properties.

Let A denote an uncertain outcome and let p denote the subjective value of P(A). Let B_r denote an outcome having known probability r. The subject is offered a sequence of choices of prospects. At any step, Prospect A is a payoff of a positive amount $g(r)$ if A occurs and Prospect B is a payoff of $g(r)$ if B_r occurs. It remains to describe the sequence $\{r_n\}$ of r-values and to choose $g(r)$. This is to be done in such a way that "honest" answers maximize the utility to the subject.

Consider the following particular choice. At Step 1, $r = r_1 = \frac{1}{2}$. At Step 2, $r = r_2 = \frac{1}{4}$ or $\frac{3}{4}$ according as Prospect B or A was chosen at Step 1. Similarly at Step n, $r = r_n = r_{n-1} - 2^{-n}$ if the preceding choice was B or $r_n = r_{n-1} + 2^{-n}$ if it was A. The idea, of course, is to obtain a sequence of r-values converging to p.

3.3 Honest and dishonest choices.

At any step n we will say the choice of A is honest if and only if $p \geq r_n$, and the choice B is honest if and only if $p \leq r_n$. To see why some choices of $g(r)$ may encourage dishonest choices, consider the fol'owing example. Suppose the procedure is to be carried to only two steps, and suppose we make the fairly natural choice of a unit payoff in each case, that is, $g(r) = 1$ for $r = 1/2, 1/4, 3/4$. Then the choices AA, AB, BA, BB have respectively the utilities 2p, p + 3/4, p + 12 and 3/4. Therefore the response AB is advantageous for all $0 \leq p \leq 3/4$ rather than only for $1/2 \leq p \leq 3/4$. If $p = 1/3$, say, a slight loss arising from a "dishonest" answer (A) at Step 1 is more than compensated for by a guaranteed utility of $3/4$ at Step 2.

Similarly another fairly natural choice of payoffs would be $g(r) = 1/r$, so that Prospect B has utility equal to unity on each occasion. This choice also provokes dishonest responses since with $p = 2/3$ we find BA has utility $11/3$ while the "honest" AB has utility $7/3$.

3.4 Characterization of payoffs which encourage honest responses.

Consider the case $n = 2$. We have:

Response	"Honest range"	Utility
BB	$0 \leq p \leq 1/4$	$(1/2)g(1/2) + (1/4)g(1/4)$
BA	$1/4 \leq p \leq 1/2$	$(1/2)g(1/2) + pg(1/4)$
AB	$1/2 \leq p \leq 3/4$	$pg(1/2) + (3/4)g(3/4)$
AA	$3/4 \leq p \leq 1$	$pg(1/2) + pg(1/4)$

We will say that the scheme is "strictly proper" (meaning that it encourages honest answers) if and only if each utility is maximum over the indicated "honest range," with ties occurring only at the boundary points between ranges. (In accordance with the terminology for score functions used in Section 2, the scheme would be called "proper" if ties were allowed over a range of values.) Since the utilities are all linear functions of p, the scheme will be strictly proper if and only if: (i) the slopes form a strictly increasing sequence in the order listed, and (ii) neighboring lines (those whose "honest ranges" meet) must intersect at the common value of p.

Condition (i) holds automatically at $p = 1/4$ and $3/4$, and at $p = 1/2$ gives

$$(3.1) \qquad g(1/2) > g(1/4).$$

Condition (ii) is also automatic at $p = 1/4$ and $3/4$, and at $p = 1/2$ we get

$$(3.2) \qquad g(3/4) = (2/3)g(1/4).$$

Similar analysis for $n = 3$ yields the following: From condition (ii) for $p = 1/4, 1/2, 3/4$ we get

$$(3.3) \qquad g(7/8) = (6/7)g(5/8),\ g(5/8) = (4/5)g(3/8),\ g(3/8) = (2/3)g(1/8)$$

and from condition (i):

$$(3.4) \qquad g(3/4) > g(5/8)$$

$$(3.5) \qquad g(1/2) > g(1/4) + g(3/8)$$

$$(3.6) \qquad g(1/4) > g(1/8).$$

If we define

$$(3.7) \qquad h(1) = 1, \text{ and } h(k) = \frac{2\cdot 4\cdots(k-3)(k-1)}{3\cdot 5\cdots(k-2)k} \quad \text{for } k = 3,5,7,\ldots,$$

then (3.3) can be written as

$$(3.8) \qquad g(k/8) = \frac{k-1}{k}\, g((k-2)/8) = h(k)g(1/8), \quad k = 3,5,7.$$

If in (3.4) we substitute from (3.2) and (3.8), we get $g(1/4) \geq (4/5)g(1/8)$, which is implied by (3.6). Also (3.1) is implied by (3.5). Thus necessary and sufficient conditions for $g(r)$ to be proper for $n = 1,2,3$ are the equalities (3.2), (3.8) and the inequalities (3.5), (3.6).

We will now give necessary and sufficient conditions for the payoffs $g(r)$ to give a strictly proper scheme for all values of n from 1 to N.

Condition (ii) is satisfied automatically at odd-numbered common values of p ($p = 2^{-n}s$, s odd). At even-numbered points, (ii) gives

$$(3.9) \qquad g(2^{-n}k) = \frac{k-1}{k} g(2^{-n}(k-2)) = h(k)g(2^{-n}), \quad k = 3,5,7,\ldots, 2^n-1.$$

Let $\beta_s^{(n)}$, $s = 1,2,\ldots, 2^n$, denote the slope of the utility of the response which is honest at time n for the range $2^{-n}(s-1) \leq p \leq 2^{-n}s$. Condition (i) can be written

$$(3.10) \qquad \beta_{s+1}^{(n)} > \beta_s^{(n)}$$

for $s = 1,2,\ldots, 2^n-1$ and $n = 1,2,\ldots, N$. For odd s, (3.10) is satisfied automatically since the difference is

$$(3.11) \qquad \beta_{s+1}^{(n)} - \beta_s^{(n)} = g(2^{-n}s)$$

which is assumed to be positive. For $s = 2,4,\ldots, 2^n-2$, define k and m by

$$(3.12) \qquad 2^{-n}s = 2^{-m}k, \quad k \text{ an odd integer.}$$

That is, $k/2^m$ equals $s/2^n$ reduced to lowest terms. Then the difference of slopes in (3.10) can be shown to be

$$(3.13) \qquad \beta_{s+1}^{(n)} - \beta_s^{(n)} = g(2^{-m}k) - \sum_{j=1}^{n-m} g(2^{-m-j}(2^jk-1)).$$

Thus if (3.9) holds then (3.10) is equivalent to

$$(3.14) \qquad g(2^{-m}) > \frac{1}{h(k)} \sum_{j=1}^{n-m} h(2^jk-1)g(2^{-m-j})$$

for $k = 1,3,\ldots, 2^m-1$, $m = 1,2,\ldots, n$, and $n = 1,2,\ldots, N$. There are many redundancies in (3.14). The inequalities in which $n = N$ imply all others. Indeed, as we show in the Appendix, the inequalities in which $k = 1$ imply all others.

Summing up, we assert that the payoffs $g(r)$ give a strictly proper scheme for $n = 1,2,\ldots, N$ if and only if (3.9) and (3.14), where (3.14) need hold only for $k = 1$, $n = N$ and $m = 1,2,\ldots, 2^N-1$. The conditions on $g(2^{-m})$ for $m = 1,2,\ldots, N$ in (3.14) are equivalent to the conditions given by the system of inequalities

$$(3.16) \qquad g(2^{-n}) > \sum_{j=1}^{\infty} h(2^j-1)g(2^{-n-j})$$

where $g(2^{-N-1})$, $g(2^{-N-2}),\ldots$ are arbitrary positive terms. Equation (3.16) holds for all n if and only if condition (ii) holds for all n.

To find a payoff schedule $g(r)$ meeting our requirements, we first note that

$$(3.17) \qquad g(2^{-n}) \geq \sum_{j=1}^{\infty} g(2^{-n-j})$$

will imply (3.16) for all n. Thus we get a strictly proper scoring rule by taking for example $g(2^{-n}) = 2^{-n}$ for $n = 1,2,\ldots$, with other values of g uniquely defined by (3.9).

3.5 <u>The limit as n tends to infinity</u>.

We will now show that if the payoffs $g(r)$ satisfy (3.9) and (3.16) for every n, then they yield a strictly proper scoring rule in the limit as n tends to infinity. Let x, y be any fixed points, $0 < x < y < 1$. We can find m, k so that k is odd and $x \leq 2^{-m}k < y$. Letting $\beta^{(m)}(z)$ denote the slope at step m at point z, we have $\beta^{(m)}(x) \leq \beta_k^{(m)}$ and $\beta^{(m)}(y) \geq \beta_{k+1}^{(m)}$. Thus by (3.11), $\beta^{(m)}(y) - \beta^{(m)}(x) \geq g(2^{-m}k)$. If we go to $n = m + m'$ steps, then $\beta^{(n)}(x) \leq \beta_\ell^{(n)}$ and $\beta^{(n)}(y) \geq \beta_{\ell+1}^{(n)}$ where

$\ell = 2^{m'}k$. By (3.13) we have

$$\beta^{(n)}(y) - \beta^{(n)}(x) \geq \beta^{(n)}_{\ell+1} - \beta^{(n)}_{\ell} \tag{3.18}$$

$$= g(2^{-m}k) - \sum_{j=1}^{m'} g(2^{-m-j}(2^j k-1))$$

$$= h(k)g(2^{-m}) - \sum_{j=1}^{m'} h(2^j k-1)g(2^{-m-j}).$$

Consider the limit of the last expression as m' (or n) tends to infinity. When $k = 1$ the limit is strictly positive whenever the strict inequalities (3.16) hold. When $k = 3, 5, \ldots$, the limit is again strictly positive by the monotone property of $h(2^j k-1)/h(k)$ proved in the Appendix. Therefore we conclude that in the limit, the scoring rule is strictly proper. Furthermore, if we let x and y approach $2^{-m}k$ from below and above, we see incidentally that the limit function is not differentiable at any of the points $2^{-m}k$.

Sequential schemes in which the score functions are strictly proper only at the finite stages $n = N$ (not necessarily proper for $n < N$) or only in the limit have been characterized by Hendrickson [9] by systems of equations and inequalities which relax the conditions (3.9) and (3.16).

4. Avoiding Assumptions about Utility Functions.

The analysis of Sections 2 and 3 is valid when the subject's utility is a linear function of the prizes or scores. There are several ways in which the procedures can be modified to avoid this assumption. One way is simply to use lottery tickets as rewards. This method applies to the procedures of Section 2 for any score function which is bounded. We simply transform the scores by a linear function of positive slope such that the transformed values lie in the interval $(0, 1)$. The subject is awarded lottery tickets such that his chance of winning the lottery equals the transformed score. The payoffs described in Section 3 will in general be bounded in total amount for any number of steps, so that they may also be replaced by parcels of lottery tickets for a single lottery. Techniques of this kind have been suggested previously by Savage [14] and Smith [17, Sec. 13].

It is also possible to employ auxiliary randomization at other points in the procedure to avoid assumptions about utilities. Here is an example. Let A be an outcome whose subjective probability p is to be determined, and let two independent auxiliary trials have outcomes $B, \bar{B}$ (where $\bar{B}$ = complement of B) and $C, \bar{C}$ whose probabilities $b = P(B)$ and $c = P(C)$ can be chosen. Let the payoff to the subject be one dollar if either AB or $\bar{B}C$ occurs, and nothing otherwise. The expectation is then $c(1-b) + bp$ (measured in utiles if the utility of one dollar is one utile). The values c and b can be chosen to depend on a stated value, say p', of $P(A)$, as we have done above in Section 2. For example, if $b = b(p') = p'/2$ and $c = c(p') = (2-p'^2)/(4-2p')$, then

$$(4.1) \qquad c(1-b) + bp = (1/4)\{2 + p^2 - (p-p')^2\},$$

which for any p is maximized by putting $p' = p$. Therefore the procedure is proper in the sense of Section 2, and avoids utilities since there is only a single prize of one dollar involved.

Proper sequential procedures for determining the subjective probability of an event A also can be made utility free by using any fixed single prize, say two utiles, and by employing independent random events. Unlike the procedure of Section 3, a single payoff is given only at time N

where N is a random variable whose value is independently determined after the sequence of choices has been made. The single payoff may be considered an advantage.

Let H have probability $1/2$. At any step let Prospect A be a payoff of two utiles if either AH or $\overline{B}_r\overline{H}$ occurs, and let Prospect B be a payoff of two utiles if either B_rH or $\overline{A}\,\overline{H}$ occurs. Here the sequence of r values is defined just as in Section 3.2, and B_r again has probability r.

Let p be the subjective probability of A, and let $q = \lim r_n$. Corresponding to the representation of $3/4$, for example, as $1/2 + 1/4 + 1/8 - 1/16 - 1/32 - \ldots$ or as $1/2 + 1/4 - 1/8 + 1/16 + 1/32 + \ldots$ there are unfortunately two honest responses when $p = 3/4$ and two corresponding sequences r_n such that $\lim r_n = 3/4$. For convenience we will assume that Prospect B is always chosen when $p = r_n$ for some step n. This is actually no restriction because it can be shown that the utilities of the two honest sequences are identical in the present scheme. With this convention every q determines a unique sequence r_n, and we see that at any step n the utility of the payoff based on the sequence r_n approaching any value q is

$$(4.2) \qquad f_n(q) = \begin{cases} p + 1 - r_n & \text{if } q > r_n \text{ (Prospect } A \text{ chosen)} \\ r_n + 1 - p & \text{if } q \leq r_n \text{ (Prospect } B \text{ chosen).} \end{cases}$$

Thus if the random variable N has distribution $P(N = n) = \pi_n$, then the utility as a function of p and q is given by

$$(4.3) \qquad E_p(f_N(q)) = \sum_{q>r_n} (p+1-r_n)\pi_n + \sum_{q\leq r_n} (r_n+1-p)\pi_n$$

$$= \alpha(q) + p\beta(q) + 1$$

where

$$(4.4)\qquad \alpha(q) = -\sum_{q>r_n} r_n\pi_n + \sum_{q\le r_n} r_n\pi_n$$

$$(4.5)\qquad \beta(q) = \sum_{q>r_n} \pi_n - \sum_{q\le r_n} \pi_n .$$

The procedure is proper if for each fixed p, the maximum of (4.3) occurs at $q = p$. It can be shown that if $r_s = 2^{-s}k$ for some odd integer k, then

$$(4.6)\qquad \alpha(r_s-) - \alpha(r_s+) = r_s(\beta(r_s+) - \beta(r_s-))$$

and

$$(4.7)\qquad \beta(r_s+) - \beta(r_s-) = 2[\pi_s - \sum_{j=s+1}^{\infty} \pi_j].$$

These are the only points of increase or decrease of α and β, and (4.6) implies that when $q = p$, (4.3) is a continuous function $H(p) = E_p(f_N(p))$. If (4.7) is nonnegative, then $H(p)$ is convex. Thus by the argument of Section 2, a necessary and sufficient condition for the procedure to be proper is

$$(4.8)\qquad \pi_s \geq \sum_{j=s+1}^{\infty} \pi_j \qquad \text{for } s = 1, 2, \ldots .$$

If also all $\pi_j > 0$, then the procedures is strictly proper. These conditions hold for example if N is the number of the Bernoulli trial on which the first success S occurs where $P(S) \geq 1/2$.

APPENDIX

For h given by (3.7) define

$$q(j, k) = h(2^j k-1)/h(k). \tag{A.1}$$

The inequalities (3.14) with $k = 1$ will imply (3.14) for $k = 3,5,7,\dots,$ provided

(A.2) $q(j, k)$ decreases as $k = 1,3,5,\dots,$ increases for each j.

For $j = 1,2,\dots,$ and $k = 1,2,\dots,$ we define

$$f(j, k) = \prod_{s=1}^{2^j} \frac{2^j k + 2s - 2}{2^j k + 2s - 1} . \tag{A.3}$$

It is straightforward to verify that

$$f(j+1, k) = f(j, 2k)f(j, 2k+2) \quad j = 1,2,\dots, k = 1,2,\dots \tag{A.4}$$

and

$$f(j, k) = \frac{k+1}{k+2} \, \frac{q(j, k+2)}{q(j, k)} \quad j = 1,2,\dots, k = 1,3,5,\dots. \tag{A.5}$$

We will now show that

$$f(j, k) < (k+1)/(k+2) \quad j = 1,2,\dots, k = 1,2,\dots, \tag{A.6}$$

which by (A.5) implies (A.2).

For $j = 1$, (A.6) follows from

$$\frac{k+2}{k+1} f(1, k) = \frac{4k^2 + 8k}{4k^2 + 8k + 3} < 1. \tag{A.7}$$

Using (A.4) and induction on j, (A.6) follows from

$$\frac{k+2}{k+1} f(j+1, k) = \frac{k+2}{k+1} f(j, 2k)f(j, 2k+2) < \frac{(k+2)(2k+1)(2k+3)}{(k+1)(2k+2)(2k+4)}$$

$$= \frac{4k^2 + 8k + 3}{4k^2 + 8k + 4} < 1.$$

REFERENCES

[1] Aczel, J. and Pfanzagl, J. (1965). Remarks on the Measurement of Subjective Probability and Information. Metrika 11 91-105.

[2] Borel, Émile (1924). A-propos d'un traité de probabilité. Revue Philosophique 98 321-336.

[3] ________ (1943). Les Probabilités et la Vie. Paris: Presses Universitaires de France.

[4] ________ (1962). Probabilities and Life. New York: Dover Publications. (English translation of [3] by Maurice Baudin.)

[5] Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review 78 1-3.

[6] De Finetti, B. (1962). Does it Make Sense to Speak of 'Good Probability Appraisers'? In Good (1962), 357-364.

[7] Good, I. J. (1952). Rational Decisions. Journal of the Royal Statistical Society, Series B 14 107-114.

[8] ________ (ed.)(1962). The Scientist Speculates. New York: Basic Books.

[9] Hendrickson, Arlo D. (1970). Payoffs to Probability Forecasters. University of Minnesota PhD Thesis and Department of Statistics Technical Report No. 145.

[10] Hendrickson, Arlo D. and Buehler, Robert J. (1971). Proper Scores for Probability Forecasters. Annals of Mathematical Statistics (to appear).

[11] Kyburg, Henry E., Jr. and Smokler, Howard E. (1964). Studies in Subjective Probability. New York: John Wiley and Sons.

[12] McCarthy, J. (1956). Measures of the Value of Information. Proceedings of the National Academy of Science. 42 654-655.

[13] Mosteller, Frederick and Wallace, David L. (1964). Inference and Disputed Authorship: The Federalist. Reading, Massachusetts; Addison-Wesley.

[14] Savage, Leonard J. (1954). The Foundations of Statistics. New York: John Wiley and Sons.

[15] ________ (1970). The Elicitation of Personal Probabilities and Expectations. Yale University. Unpublished.

[16] Shuford, E. H., Jr., Albert, A., and Massengill, H. E. (1966). Admissible Probability Measurement Procedures. Psychometrika 31 125-145.

[17] Smith, Cedric A. B. (1961). Consistency in Statistical Inference and Decision. Journal of the Royal Statistical Society, Series B 23 1-37.

[18] Staël von Holstein, Carl-Axel S. (1970). Assessment and Evaluation of Subjective Probability Distributions. The Economic Research Institute at the Stockholm School of Economics, Stockholm.

[19] Winkler, Robert L. (1967a). The Quantification of Judgement: Some Methodological Suggestions. Journal of the American Statistical Association 62 1105-1120.

[20] ________ (1967b). The Quantification of Judgement: Some Experimental Results. Proceedings of the Business and Economic Statistics Section, American Statistical Association 386-395.

[21] ________ (1969). Scoring Rules and the Evaluation of Probability Assessors. Journal of the American Statistical Association 64 1073-1077.

[22] Winkler, Robert L. and Murphy, Allan H. (1968). 'Good' Probability Assessors. Journal of Applied Meteorology 7 751-758.